UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS

Submitted by:

Maulana Arya Alambana
22/492322/PA/21104

With the increasing use of Large Language Models (LLMs) for code generation, several challenges arise in verifying code authenticity, particularly in education and the software security industry. Currently existing detection tools, which often rely on probabilistic and classification-based methods, often struggle with generalization across programming styles. This study introduces a zero-shot synthetic code detection framework that utilizes CodeT5 as the embedding model, leveraging its encoder–decoder architecture to better preserve semantic, syntactic, and stylistic information during rewriting, aiming to capture the behavioral difference between human-written and machine-generated code. Using cosine similarity between original and rewritten Python code, the method is evaluated on a stylistically diverse dataset designed to stress-test generalization across prompting conditions, and its performance is compared against representations derived from GraphCodeBERT to assess the role of model architecture in rewriting-based detection. Results show that all four configurations exceed the random guessing baseline, with CodeT5-based models achieving a higher average AUROC (0.5805) compared to GraphCodeBERT-based models (0.5731), though detection performance varies substantially across dataset variants, with structural modifications such as identifier renaming reducing effectiveness to near-random levels.

Keywords: Synthetic Code Detection, Large Language Models, CodeT5, Rewriting Similarity, Zero-Shot Learning.

UNIVERSITAS
GADJAH MADA
LARGE LANGUAGE MODEL-GENERATED SYNTHETIC CODE DETECTION VIA REWRITING-BASED COMPARATIVE ANALYSIS
Maulana Arya Alambana, Arif Nurwidyantoro, S.Kom., M.Cs., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

DETEKSI KODE SINTETIS YANG DIHASILKAN OLEH MODEL BAHASA BERSKALA BESAR MELALUI ANALISIS KOMPARATIF BERBASIS PENULISAN ULANG

Oleh:

Maulana Arya Alambana
22/492322/PA/21104

Dengan meningkatnya penggunaan Large Language Models (LLMs) untuk generasi kode, muncul berbagai tantangan dalam memverifikasi keaslian kode, khususnya di bidang pendidikan dan keamanan perangkat lunak. Alat deteksi yang ada saat ini umumnya bergantung pada metode probabilistik dan klasifikasi yang kesulitan melakukan generalisasi terhadap beragam gaya pemrograman. Penelitian ini memperkenalkan kerangka kerja zero-shot synthetic code detection berbasis rewriting menggunakan CodeT5 sebagai model embedding, yang memanfaatkan arsitektur encoder–decoder untuk menangkap perbedaan perilaku rewriting antara kode buatan manusia dan kode yang dihasilkan LLM, dievaluasi pada dataset beragam gaya dari Pan et al. (2024) dan dibandingkan dengan GraphCodeBERT. Hasil menunjukkan bahwa keempat konfigurasi model melampaui baseline tebakan acak, dengan model berbasis CodeT5 mencapai rata-rata AUROC lebih tinggi (0,5805) dibandingkan dengan GraphCodeBERT (0,5731), meskipun modifikasi struktural seperti penggantian nama identifier menurunkan performa hingga mendekati tingkat acak.

Kata Kunci: Deteksi Kode Sintetis, Model Bahasa Berskala Besar, CodeT5, Kesamaan Penulisan Ulang, Pembelajaran Zero-Shot.